

Hybrid Multi-Task NLP for Comprehensive Mental Health Assessment

Vishal Agarwal
Dept. of Computer Science and
Engineering (AI)
Manipal Institute of Technology
Bengaluru, India
vishal.mitblr2023@learner.manipal.edu

Abstract—The rising prevalence of mental health disorders in the modern era—driven by factors such as genetic vulnerability, early-life trauma, social isolation, economic inequality, and the pressures of urban living—underscores the need for scalable and intelligent assessment tools. Natural Language Processing (NLP) has emerged as a transformative technology in healthcare, enabling the analysis of user-generated text to detect psychological distress. This paper presents MindScan, a hybrid, multi-task NLP system designed to perform mental health classification, emotion recognition, and sentiment analysis from a single input. By integrating Transformer-based models with Parameter-Efficient Fine-Tuning (PEFT) and class-weighted loss optimization, the system achieves high contextual sensitivity and interpretability.

Keywords—Hybrid Architecture, Multi-Task Learning, Psychological State Classification, Emotion Recognition, Sentiment Analysis, Transformer Models, PEFT

I. INTRODUCTION

The rising prevalence of mental health disorders in the modern era—shaped by factors such as genetic vulnerability, early-life trauma, social isolation, economic inequality, and the pressures of fast-paced urban living—highlights an urgent global health concern. This calls for the development of scalable and intelligent tools for assessment and intervention. In recent years, Machine Learning (ML) and Natural Language Processing (NLP) have emerged as transformative technologies in healthcare, offering sophisticated algorithms capable of interpreting unstructured data, predicting outcomes, and supporting clinical decision-making. Prior research has demonstrated the effectiveness of NLP in analysing user-generated text to detect psychological distress, including depression, anxiety, and suicidal ideation [1].

Mental health is not merely the absence of illness—it profoundly influences how individuals think, feel, and act, shaping how they manage stress, relate to others, and make decisions [4]. Many individuals struggle to articulate or recognize their emotions and experiences. Consequently, NLP systems must be capable of discerning subtle indicators and latent psychological patterns embedded within natural language, requiring models that can handle ambiguity, context, and the nuanced semantics of human expression.

Recent global reports emphasize the urgency of this challenge. In 2025, over one billion people worldwide were reported to be living with a mental disorder, with anxiety

and depression being the most prevalent [2]. Furthermore, the proportion of individuals likely to suffer from anxiety reached a record high of 23%, up from 17% in 2022, indicating a worsening trend in emotional stress indicators [3].

This paper presents *MindScan*, a hybrid, multi-task NLP system designed for comprehensive psychological and emotional assessment from a single text input. Unlike systems focused solely on mental health classification, MindScan performs three distinct yet complementary tasks:

- **Mental health condition classification** using a fine-tuned RoBERTa model, outputting probabilities across seven categories: Normal, Depression, Suicidal, Anxiety, Stress, and Bipolar and Personality Disorder
- **Emotion detection** using a pretrained DistilBERT model further optimized via Parameter-Efficient Fine-Tuning (PEFT) with a LoRA configuration, identifying nuanced emotional states such as joy, sadness, anger, fear, love, and surprise
- **Sentiment analysis** using a second fine-tuned DistilBERT model to determine whether the overall tone of the input is positive or negative

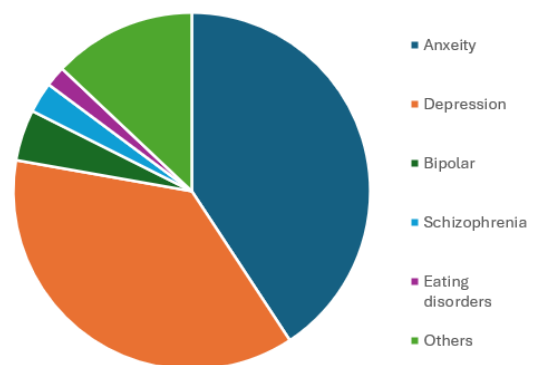


Figure 1: Global distribution of mental health conditions based on 2021 prevalence data. (as reported by WHO) [8]

By integrating Transformer-based deep learning with a multi-task design, the system achieves high interpretability, contextual sensitivity, and multi-dimensional insight into the user's psychological state. This architecture ensures that

MindScan provides rich, layered assessments rather than simplistic, single-label predictions.

The remainder of this paper is structured as follows: Section II reviews related work, focusing on ML in healthcare and NLP applications for psychological and emotional assessment. Section III details the proposed hybrid methodology, including data preparation and model fine-tuning procedures. Section IV presents the experimental setup, results, and performance analysis across all three classification tasks. Finally, Section V concludes the paper by summarizing the contributions and outlining avenues for future research.

II. LITERATURE SURVEY

The intersection of Natural Language Processing (NLP) and mental health has gained significant attention in recent years, with researchers exploring how linguistic patterns can reveal psychological states. Early work by Calvo et al. [1] demonstrated that non-clinical text sources such as social media and online forums could be mined using NLP techniques to detect signs of depression, anxiety, and suicidal ideation. Their findings laid the groundwork for using language as a proxy for psychological evaluation.

Traditional machine learning models have long been employed for mental health classification tasks. Resnik et al. [5] used TF-IDF features with classifiers like Logistic Regression and Support Vector Machines to identify depressive language in Reddit posts. These approaches offered interpretability and computational efficiency, but often lacked the contextual depth required to capture nuanced psychological signals.

With the rise of deep learning, Transformer-based models such as BERT and its variants have become the state-of-the-art in NLP. These models capture contextual dependencies and semantic nuances more effectively than shallow methods. The GoEmotions dataset introduced by Demszky et al. [6] enabled fine-grained emotion classification using BERT-based architectures, covering 27 emotion labels and sentiment polarity. Their work demonstrated that pretrained language models could be fine-tuned to detect subtle emotional cues embedded in everyday language.

Multi-task learning has also emerged as a powerful paradigm in psychological NLP. Zhang et al. [7] proposed a multi-task framework that jointly learns emotion and sentiment classification, showing that shared representations across tasks improve generalization and reduce overfitting. Such architectures are particularly relevant when analysing complex psychological states that span multiple dimensions.

While prior studies have explored mental health classification, emotion detection, and sentiment analysis as separate tasks, few systems have attempted to unify these dimensions into a cohesive framework. Most existing systems focus narrowly on either psychological condition identification or emotional tone, often overlooking the nuanced interdependencies between them.

In this system, Transformer-based models are fine-tuned for each task, including Parameter-Efficient Fine-Tuning (PEFT) with LoRA for emotion detection. This layered approach enables more holistic interpretation of linguistic cues, offering richer insights into the user's mental and emotional state than single-task models typically provide.

AUTHOR	TASK	MODEL USED	DATASET	LIMITATION
Calvo et al. (2017)	Mental health detection from text	Rule-based + ML classifiers	Non-clinical text sources	Limited scalability; lacks deep contextual modelling
Resnik et al. (2015)	Depression classification	TF-IDF + Logistic Regression	Reddit posts	Shallow features; limited to binary classification
Demszky et al. (2020)	Emotion classification	BERT	GoEmotions	Focused only on emotions; no mental health or sentiment integration
Zhang & Wang (2020)	Multi-task emotion & sentiment	Shared Transformer encoder	Custom multi-label corpus	No psychological condition modelling; limited interpretability

Table 1: Summary of prior works in psychological NLP, highlighting task focus, model architecture, datasets used, and limitations.

III. METHADODOLOGY

This section describes the design, implementation, and training of the proposed hybrid multi-task NLP system for psychological and emotional assessment. The system processes a single user-generated text input and performs three classification tasks: mental health condition identification, emotion recognition, and sentiment analysis. Each task is handled by a dedicated Transformer-based model, fine-tuned for its respective objective.

A. System Architecture Overview

The system adopts a modular multi-task architecture, where each classification task is handled by an independently fine-tuned model. This design ensures task-specific optimization while maintaining flexibility and interpretability.

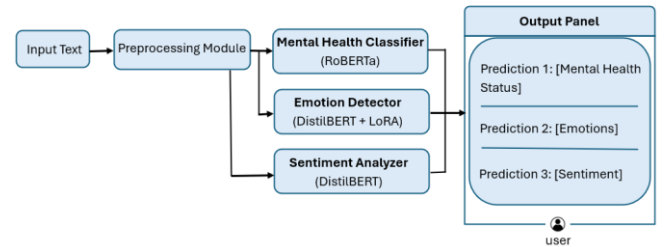


Figure 2: System architecture of the proposed multi-task NLP framework for psychological and emotional assessment. (The outputs are displayed to the user via an integrated output panel)

B. Datasets Used

Two publicly available datasets were used to train and evaluate the system, with distinct files selected for each classification task

- **Mental Health Condition Classification:** The Sentiment Analysis for Mental Health dataset [12] was used to train the mental health classifier. It contains labelled text samples across psychological categories such as Normal, Depression, Suicidal, Anxiety, Stress, and Bipolar/Personality Disorder. The dataset was curated from diverse platforms including social media posts, Reddit posts, Twitter posts, other Kaggle datasets, and more.
- **Emotion Recognition:** The Sentiment and Emotion Analysis Dataset [11] was used for both emotion detection and sentiment classification. For emotion detection, the `combined_emotion.csv` file was reduced to 29,997 samples using stratified sampling to preserve class balance across six emotion categories: Joy, Sadness, Anger, Fear, Love, and Surprise. This subset was split into 26,997 training samples and 3,000 evaluation samples
- **Sentiment Analysis:** The same dataset [11] also provided the `combined_sentiment_data.csv` file, which was used to train the sentiment analyser. This file contains binary sentiment labels (Positive and Negative) mapped to user-generated text.

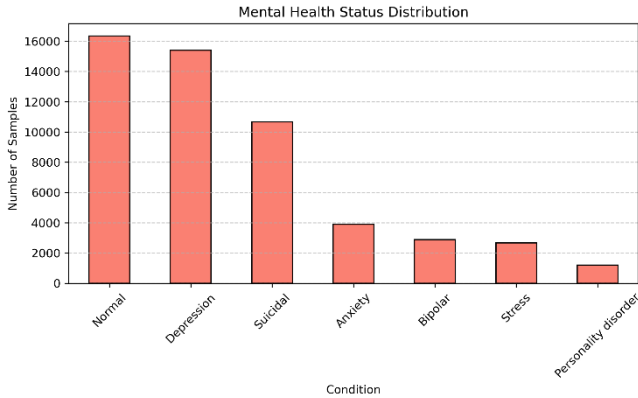


Figure 3(a): Label distribution in the full mental health dataset.

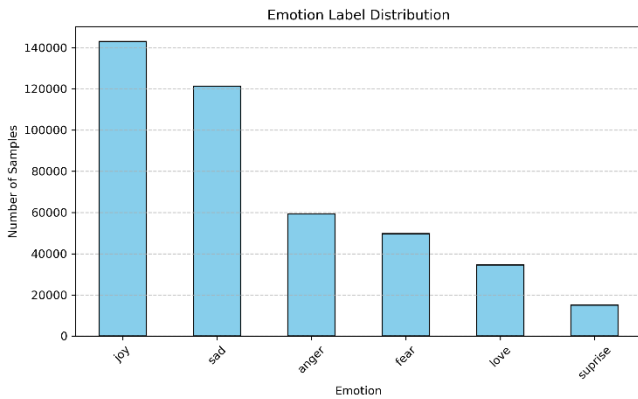


Figure 3(b): Label distribution in the full `combined_emotion.csv` file of the sentiment and emotion analysis dataset (prior to sampling due to the computational constraints)

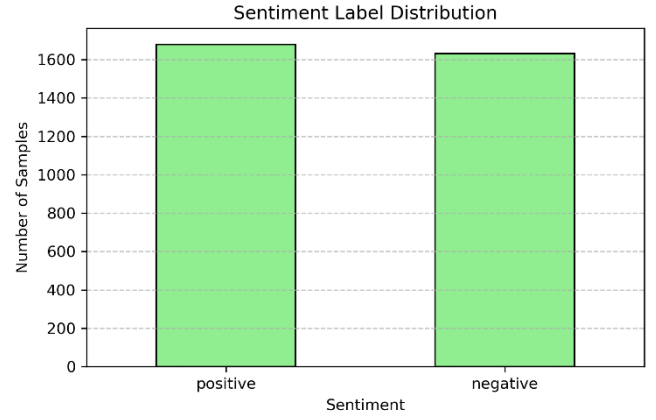


Figure 3(c): Label distribution in the full `combined_sentiment_data.csv` file of the sentiment and emotion analysis dataset

C. Preprocessing Pipeline And Model Configuration

1. Mental Health Status Classification

- Removed entries with missing or empty statement or status fields
- Stripped whitespace and converted all statements to string format
- Removed duplicate statements to ensure unique samples
- Mapped six psychological categories (Normal, Depression, Suicidal, Anxiety, Stress, Bipolar/Personality Disorder) to integer labels
- Applied stratified train-test split (90/10) to preserve class balance
- Computed class weights using sklearn's `class_weight` utility to address label imbalance
- Tokenized text using RoBERTa tokenizer with truncation and padding to a maximum sequence length of 128 tokens
- Retained only input `ids`, `attention_mask`, and label columns for training
- Converted datasets to PyTorch format for compatibility with Hugging Face Trainer

It was built using a custom-weighted RoBERTa model. A wrapper around `RobertaForSequenceClassification` was implemented to incorporate class-weighted cross-entropy loss, improving robustness against label imbalance. The model predicts one of six psychological conditions using a SoftMax activation. Tokenization was performed using Hugging Face's `AutoTokenizer`, and the entire pipeline was implemented using Transformers and PyTorch [10].

2. Emotion Recognition

- Lowercased emotion labels and stripped whitespace
- Corrected label typos (e.g., 'sad' → 'sadness', 'suprise' → 'surprise')
- Mapped six emotion classes (Sadness, Joy, Love, Anger, Fear, Surprise) to integer labels
- Applied stratified sampling to reduce the dataset to 30,000 samples for CPU stability
- Split the dataset into 90% training and 10% evaluation using a fixed seed (42)

- Tokenized text using DistilBertTokenizerFast with truncation and padding to a maximum sequence length of 128 tokens
- Retained only input_ids, attention_mask, and label columns for training
- Converted datasets to PyTorch format for compatibility with Hugging Face Trainer

It was built using the pretrained bhadesh-savani/distilbert-base-uncased-emotion [9] as the backbone. To enable efficient fine-tuning, Low-Rank Adaptation (LoRA) was applied using the PEFT framework. The LoRA configuration targeted the q_lin and v_lin layers of the attention mechanism with rank $r = 8$, scaling factor $\alpha = 16$, and dropout 0.1. The output layer predicts one of six emotion classes using a SoftMax activation. Training was guided by the cross-entropy loss function. Tokenization was performed using DistilBertTokenizerFast, and the entire pipeline was implemented using Hugging Face Transformers, PEFT, and PyTorch.

3. Sentiment Analysis

- Mapped binary sentiment labels (Positive, Negative) to integer values
- Converted the dataset to Hugging Face Dataset format
- Applied an 80/20 train-test split using a fixed seed (42) for reproducibility
- Tokenized text using DistilBertTokenizerFast with truncation and padding to a maximum sequence length of 128 tokens
- Retained only input_ids, attention_mask, and label columns for training
- Converted datasets to PyTorch format for compatibility with Hugging Face Trainer

It was built using the distilbert-base-uncased model as the backbone. The architecture consists of a DistilBERT encoder followed by a binary classification head. The output layer uses a SoftMax activation to predict one of two sentiment classes. Model training was guided by the cross-entropy loss function, suitable for binary classification tasks. Tokenization was performed using Hugging Face’s DistilBertTokenizerFast, and the entire pipeline was implemented using Hugging Face Transformers and PyTorch.

PARAMTER	MENTAL HEALTH CLASSIFIER	EMOTION DETECTOR	SENTIMENT ANALYSER
Base Model	RoBERTa-base	Pre-trained Distilbert model	Distilbert-base
Epochs	3	5	3
Train Batch Size	32	8	16
Eval Batch Size	64	64	64
Gradient Accumulation	4	4	-
Learning Rate	2e-5	5e-5	3e-5
Weight Decay	0.1	0.1	0.05

Optimizer	AdamW	AdamW	AdamW
Evaluation Strategy	Epoch-end	Epoch-end	Epoch-end
Save Strategy	Steps	Epoch-end	Epoch-end
Save Steps	250	-	-
Early Stopping	Patience=2	Patience=2	Not used
Metric for Best Model	Validation Loss	Validation Loss	Validation Loss
Mixed Precision (fp16)	Disabled	Disabled	Disabled
Logging	Every 100 steps	Every 100 steps	Every 100 steps
Loss Function	Weighted Cross-Entropy	Cross-Entropy	Cross-Entropy
Frameworks Used	Transformers, PyTorch	Transformers, PEFT, PyTorch	Transformers, PyTorch

Table 2: Summary of training configurations for all three models, including hyperparameters, evaluation strategies, and framework choices.

IV. RESULTS AND DISCUSSION

This section presents the evaluation results for three classification models: Mental Health, Sentiment, and Emotion. Each model was assessed using standard metrics including accuracy, macro-averaged F1-score, precision, recall, and confusion matrices. For the binary sentiment classifier, a ROC curve was also included to visualize discriminative performance.

A. Evaluation Metrics

1. Mental Health Status Classification

It was evaluated on a test set of 5,108 samples spanning seven diagnostic categories. The model achieved an accuracy of 82.69% and a macro F1-score of 0.799, indicating balanced performance across both dominant and minority classes. The macro precision and recall were 78.37% and 82.28%, respectively, suggesting that the model is slightly more sensitive than precise — a desirable trait in mental health screening where false negatives can be more critical than false positives.

The classification report (Appendix A1) reveals strong performance for dominant classes such as *Normal* (F1 = 0.96) and *Anxiety* (F1 = 0.88), while minority classes like *Personality Disorder* (F1 = 0.65) and *Suicidal* (F1 = 0.74) showed lower scores. This disparity is likely due to class imbalance, as seen in the support values, and highlights the need for targeted data augmentation or re-weighting strategies.

Notably, Depression and Suicidal were frequently confused, which aligns with their semantic and symptomatic overlap in clinical contexts. Similarly, Stress was often misclassified as Anxiety, suggesting that the model may benefit from more nuanced feature representations or context-aware embeddings.

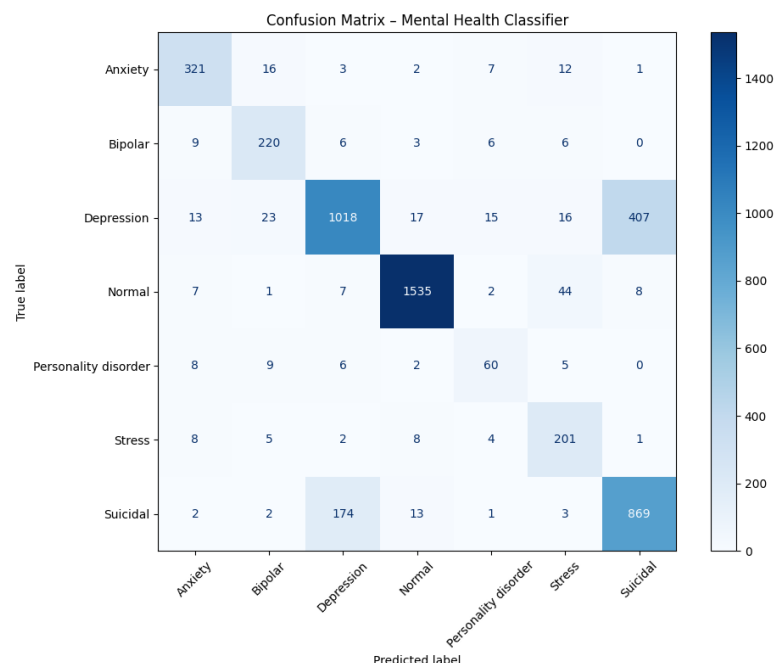


Figure 4(a): Confusion Matrix which illustrates the distribution of predicted versus actual labels across seven mental health categories.

2. Emotion Recognition

It was evaluated on a test set of 3,000 samples spanning six emotion categories. The model achieved an accuracy of 92.87% and a macro F1-score of 0.896, indicating strong and balanced performance across both high-frequency and low-frequency emotions. The macro precision and recall were 89.89% and 89.28%, respectively, suggesting that the model maintains consistent sensitivity and precision across classes - a valuable trait in emotion recognition where subtle distinctions can impact downstream applications.

The classification report (Appendix A3) reveals excellent performance for dominant emotions such as Sadness (F1 = 0.966) and Joy (F1 = 0.944), while less frequent emotions like Surprise (F1 = 0.792) and Love (F1 = 0.829) showed comparatively lower scores. This variation is likely influenced by class imbalance and the nuanced nature of emotional expression, particularly in short text formats.

Notably, Fear and Sadness were occasionally confused, which aligns with their semantic proximity and overlapping linguistic cues. Similarly, Surprise exhibited lower recall, suggesting that the model may benefit from richer contextual embeddings or multimodal cues to better capture transient emotional states.

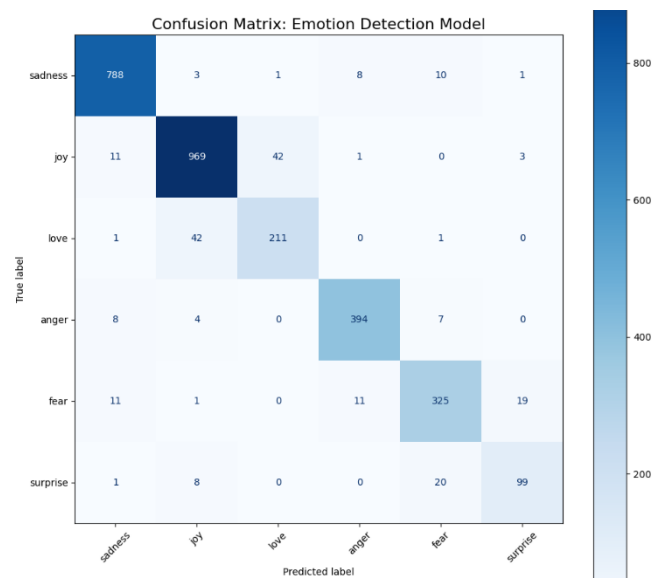


Figure 4(b): The confusion matrix which illustrates the distribution of predicted versus actual labels across six emotion categories.

3. Sentiment Analysis

The sentiment classifier was evaluated on a test set of 662 samples for binary polarity detection. It achieved an accuracy of 91.39% and a macro F1-score of 0.919, indicating high reliability in distinguishing between positive and negative sentiment. The macro precision and recall were 91.29% and 92.59%, respectively, suggesting that the model is slightly more sensitive than precise — a favourable trait in sentiment analysis, where capturing positive feedback is often prioritized in applications such as customer reviews or mental health screening.

The classification report (Appendix A2) shows balanced performance across both classes, with Positive sentiment achieving an F1-score of 0.92 and Negative sentiment scoring 0.91. The confusion matrix confirms this, revealing a clear separation between positive and negative predictions with minimal misclassification. Additionally, the ROC curve demonstrates strong discriminative power, with an AUC approaching 1.0 — indicating that the model consistently ranks positive samples higher than negative ones.

The model demonstrates high precision and recalls for both classes, with minimal misclassification. This supports the reliability of the sentiment predictions.

To further assess the model's discriminative ability, a Receiver Operating Characteristic (ROC) curve was plotted. It visualizes the trade-off between true positive rate and false positive rate across different thresholds. A curve that approaches the top-left corner indicates strong classification performance. In this case, the ROC curve shows an AUC close to 1.0, confirming the model's ability to rank positive samples higher than negative ones with high confidence.

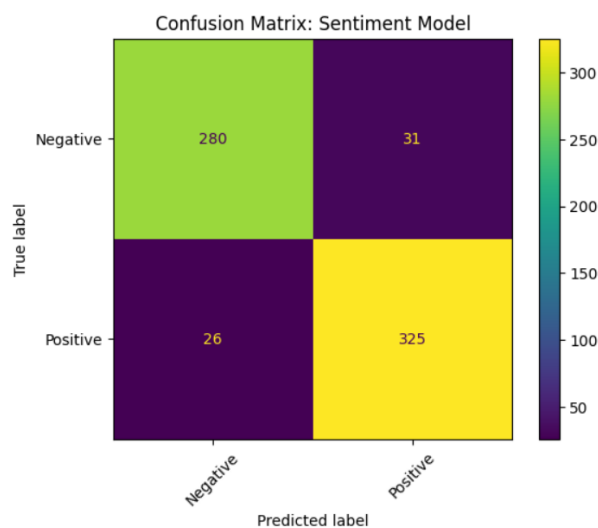


Figure 4(c): Confusion matrix which shows binary classification performance for positive and negative sentiment.

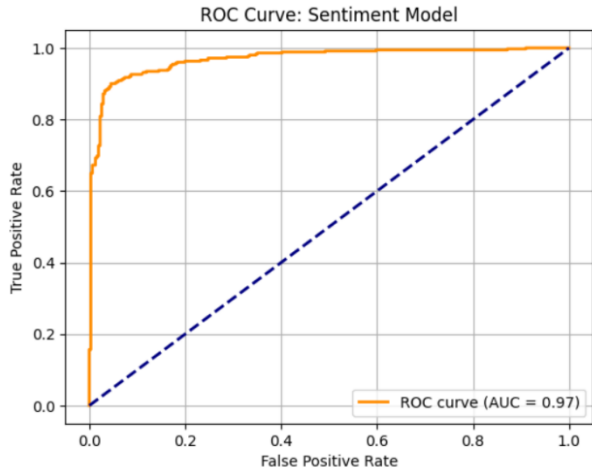


Figure 4(d): The ROC curve illustrates the model’s ability to distinguish between positive and negative sentiment.

MODEL	ACCURACY (%)	MACRO PRECISION (%)	MACRO RECALL (%)	MACRO F1-SCORE (%)
Mental Health Classifier	82.69	78.37	82.28	79.92
Emotion Classifier	92.87	89.89	89.28	89.58
Sentiment Classifier	91.39	91.29	92.59	91.94

Table 3: Comparative macro-averaged evaluation metrics across the models

B. Confidence Visualization

To enhance interpretability, model confidence distributions were visualized using a Streamlit interface. These plots display the probability scores assigned to each predicted class, helping identify borderline cases and potential annotation noise. For instance, predictions with low confidence may indicate ambiguous input or overlapping class boundaries.

This visualization aids in understanding model behaviour beyond raw metrics and supports error analysis, especially for sensitive applications like mental health screening. It also provides a practical layer of transparency for end-users interacting with the deployed model.

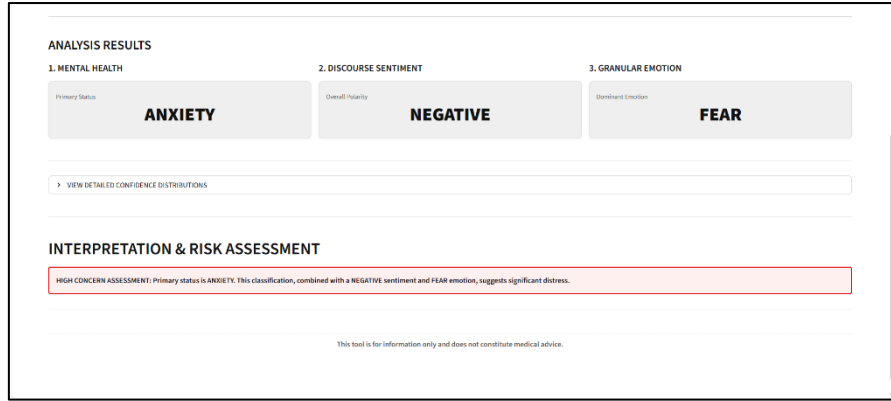


Figure 5(b): The output screen displays predictions from three models: Mental Health (Anxiety), Sentiment (Negative), and Emotion (Fear). The interpretation panel provides a high-concern risk assessment based on the combined outputs, indicating significant emotional distress.

C. Discussion

The multi-task architecture presented in this work demonstrates the practical viability of deploying lightweight Transformer models for psychological and emotional assessment. Beyond raw performance, the integration of confidence visualization and interpretability tools adds a critical layer of transparency, especially important in sensitive domains like mental health. The Streamlit interface enables real-time feedback and user accessibility, bridging the gap between technical capability and practical deployment.

While the models are effective in their respective tasks, their limitations highlight opportunities for refinement. The mental health classifier, for instance, could benefit from strategies that enhance sensitivity to underrepresented conditions. Emotion detection, being inherently nuanced, may require richer contextual modeling to capture subtle affective states. The confidence plots reveal areas of uncertainty that could be addressed through calibration or ensemble learning.

Future Work may explore:

- Data augmentation for underrepresented classes
- Context-aware embeddings for emotion detection
- Calibration techniques to improve confidence reliability
- Ensemble methods to boost minority class sensitivity

As this system moves closer to real-world deployment, ethical considerations such as user privacy, informed consent, and responsible interpretation of predictions must be prioritized. Designed for informational use only, the tool should not replace professional diagnosis or care. All datasets are publicly available, and model configurations are detailed in Table 2, with the training pipeline implemented using Hugging Face Transformers and PyTorch for full reproducibility.

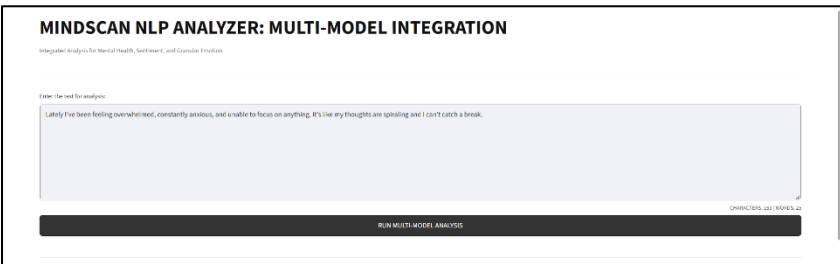
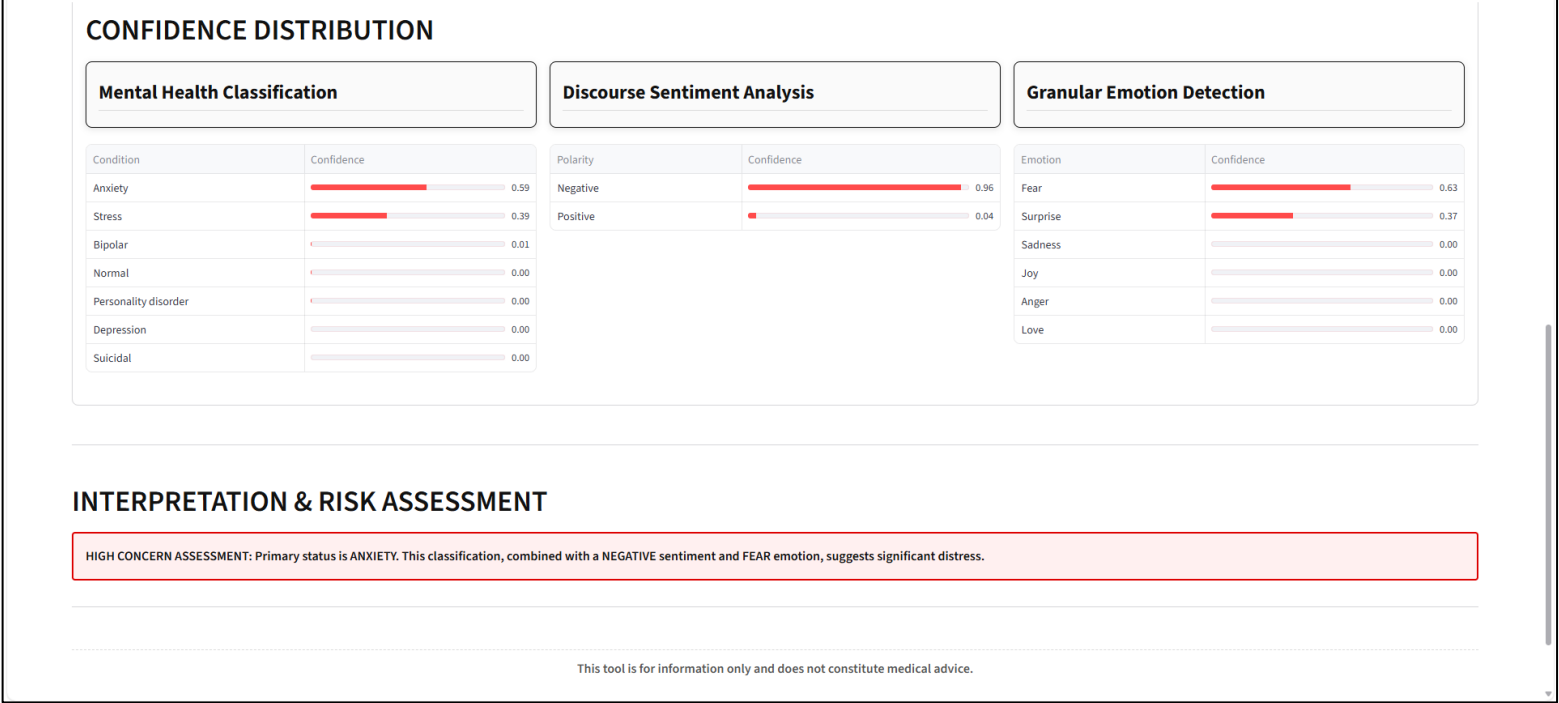


Figure 5(a): The input interface populated with a sample user entry expressing emotional distress, the analyzer accepts natural language input and displays character and word counts prior to analysis.



- [2] World Health Organization, "Over a billion people living with mental health conditions," WHO, Geneva, Switzerland, Sep. 2025. [Online]. Available: <https://www.who.int/news/item/02-09-2025-over-a-billion-people-living-with-mental-health-conditions-services-require-urgent-scale-up> [Accessed: Oct. 14, 2025].
- [3] AXA Mind Health Report, "Mental health continues to deteriorate around the world," IPSOS, 2025. [Online]. Available: <https://www.ipsos.com/en/axa-mind-health-report-mental-health-continues-deteriorate-around-world> [Accessed: Oct. 14, 2025].
- [4] Cleveland Clinic, "Why Mental Health Is So Important," 2025. [Online]. Available: <https://health.clevelandclinic.org/why-mental-health-is-so-important> [Accessed: Oct. 14, 2025].
- [5] P. Resnik, W. Armstrong, L. Claudino, T. Nguyen, V. A. Nguyen, and J. Boyd-Graber, "Beyond LDA: Exploring supervised topic modeling for depression-related language in Twitter," in *Proc. 2nd Workshop Computational Linguistics and Clinical Psychology*, 2015, pp. 1-10.
- [6] D. Demszky, D. Movshovitz-Attias, J. Ko, A. Cowen, G. Nemade, and S. Ravi, "GoEmotions: A dataset of fine-grained emotions," *arXiv preprint arXiv:2005.00547*, 2020.
- [7] Y. Zhang and D. Wang, "Multi-task learning for emotion and sentiment classification with shared representations," *IEEE Access*, vol. 8, pp. 23030–23039, 2020.
- [8] World Health Organization, *World mental health today: latest data*. Geneva: WHO, 2025. [Licence: CC BY-NC-SA 3.0 IGO].
- [9] T. Savani, "distilbert-base-uncased-emotion," Hugging Face, 2021. [Online]. Available: <https://huggingface.co/bhadresh-savani/distilbert-base-uncased-emotion> [Accessed: Oct. 14, 2025].
- [10] B. Wolf et al., "Transformers: State-of-the-art natural language processing," in *Proc. 2020 Conf. Empirical Methods Natural Language Processing: System Demonstrations*, 2020, pp. 38–45. [Online]. Available: <https://huggingface.co/transformers>
- [11] K. Kushagra, "Sentiment and Emotion Analysis Dataset," Kaggle, 2023. [Online]. Available: <https://www.kaggle.com/datasets/kushagra3204/sentiment-and-emotion-analysis-dataset> [Accessed: Oct. 14, 2025].
- [12] S. Sarkar, "Sentiment Analysis for Mental Health," Kaggle, 2023. [Online]. Available: <https://www.kaggle.com/datasets/suchintikasarkar/sentiment-analysis-for-mental-health> [Accessed: Oct. 14, 2025].
- [13] Y. Liu et al., "RoBERTa: A Robustly Optimized BERT Pretraining Approach," *arXiv preprint arXiv:1907.11692*, 2019.